Toward an abstract Wikipedia

Denny Vrandečić

Google
vrandecic@google.com

Abstract. Description logics are a powerful and widely used family of languages that allow to express ontologies and knowledge bases in a way that allows for decidable algorithms and a well-understood reasoning complexity. Ontologies formalize the shared understanding of a domain. But the most expressive and widespread languages that we know of are human natural languages, and the largest knowledge base we have is the wealth of text written in human languages.
We need a path forward to bridge the gap between formal knowledge representation languages and human natural languages. We propose a project that will simultaneously expose that gap, provide a place to constructively and collaboratively close that gap, and demonstrate the progress as we move forward: a multilingual Wikipedia.
Imagine Wikipedia not being written in a natural language, but in an abstract language which gets translated into one of its roughly 300 languages whenever someone wants to read part of it. This would make current Wikipedia editors about 100x more productive, increase the content of Wikipedia by at least 10x, probably increase the number of Wikipedians, and make the Web much more useful for many people who currently have no content interesting for them because they speak a language not widely used on the Web today. For billions of users, this will unlock knowledge that they currently do not have access to. For many Wikipedia language editions this will be the only viable chance to succeed in their mission. It is an important step toward enabling everyone to share in the sum of all knowledge.

Keywords: Semantics · Multilingual · Abstract language.

1 Wikipedia today

Wikipedia is one of the most important sources of knowledge today. Following its vision to allow everyone to share in the sum of all knowledge, it aims to create comprehensive and constantly up-to-date encyclopedias that anyone, anywhere can read and contribute to.

In order for everyone to be able to read Wikipedia, it has to be available in a language that they speak. Currently there are Wikipedia projects in about 300 languages, totaling close to 50 million articles.

But the languages are very unevenly distributed: at the time of writing, English has more than 5.6 million articles, another dozen language editions have

more than a million articles. 59 language editions have more than 100,000, and 137 languages more than 10,000 articles. Many language editions are tiny: Zulu has 901, Swati 438, Samoan 797, Cherokee 834, and Cree 131 articles.

But not only the number of articles vary widely, also the comprehensiveness of articles about the same topic over different languages can be dramatically different: whereas the German Wikipedia has a comprehensive article on Frankfurt, the Corsican Wikipedia has merely a single line of text – and a link to the Frankfurt football club. And this is not always a function of the overall size of a given Wikipedia: whereas the English Wikipedia has a single line on the Port of Călăraşi, the Romanian Wikipedia offers several paragraphs, including a picture. "Local" Wikipedias often have information that is missing from the big Wikipedias.

Besides the coverage in the articles, also the topics of the articles are vastly different: the two most active Wikipedias are the English and German language editions. The English Wikipedia has 5.6 million articles, the German Wikipedia 2.1 million – but only 1.1. Million, roughly half, of the topics of the German Wikipedia articles have also an article in the English Wikipedia. More than half a million topics are entirely missing from the English Wikipedia. Only 100,000 topics are covered by all of the top ten most active Wikipedias (whereas all of them have about a million articles or more), only 25,000 by the top 20.

There are currently Wikipedia articles about 17.9 million topics. Having them available in all 300 languages of Wikipedia would lead to more than 5.3 billion articles. This dwarfs the number of 50 million available articles - less than a percent of that goal. And this does not take into consideration the effort required to maintain these articles.

Achieving this goal is ambitious. The Wikipedia communities have currently about 69,000 active members. 31,000 of those are active on the English Wikipedia, 5,500 on German, 5,000 on French, 4,000 on Spanish. In total, eleven Wikipedias have more than a thousand active editors. The shared multimedia site Wikimedia Commons has 7,000 active contributors, the shared structured knowledge base Wikidata 8,000 active contributors. More than half of the Wikipedia language editions have less than ten active editors.

Completing and maintaining an encyclopedia with ten active volunteer editors is ambitious.

2 A multilingual Wikipedia

The underlying issue is that the size of the problem is essentially the number of topics *multiplied* with the number of languages. In the following we suggest a solution that reduces it into a problem where the size is essential the number of topics *added* to the number of languages.

We sketch the following solution (see Fig. 1): the multilingual Wikipedia consists of three main components: Content, Renderers, and Constructors.

Content. The largest component is the Content: each individual article is represented by a formal knowledge base that captures the content of an item in

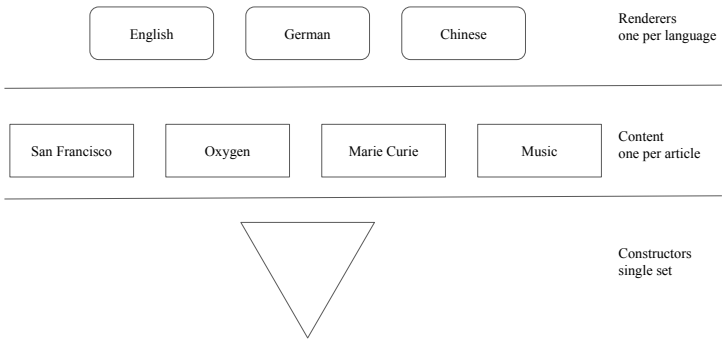

Fig. 1. The three main components of the multilingual Wikipedia are the Renderers, the Content, and the Constructors. Additional components (not pictured) include a database of lexical knowledge and a database of structured world knowledge, which the other components have access to.

an abstract, language-independent manner. We imagine that this is not just declarative about the topic of the article, but also and in particular captures e.g. the order the text is rendered in, where paragraphs and sections start and end, redundancies in the text, additional knowledge about other topics which is beneficial to the understanding of the text, etc. This is all in contrast what a traditional RDF knowledge base would hold.

Constructors. The smallest component are the Constructors: this defines the 'language' in which the individual articles in the Content are expressed. If you imagine the Content be expressed in a series of function calls, then these are the definitions of the individual functions. If you imagine the Content be expressed as Frame instantiations, then these are the definitions of the available Frames, their slots, etc. If you imagine the Content akin to the ABox of the ontology, than this is akin to the TBox.

Renderers. The Renderers have the task of translating the Content to natural language text suitable for consumption by the readers. In order to do so they need to define for each Constructor used how to represent that Constructor in natural language. In order to do so they have access to a large ontological and lexicographical knowledge base in Wikidata.

It is expected that in this system, the largest number of contributors will be working on the Content (due to its sheer size and need to keep it constantly updated). The hope is that each of the Renderers for a given language requires a small number of active contributors – maybe ten or less – in order to achieve sufficient coverage (note that the lexicographical knowledge can be maintained by

a larger crowd on Wikidata, which would reduce the workload for each individual Renderer). This way even such a small number of contributors would be enabled to create a comprehensive and up-to-date encyclopedic resource in their own language.

2.1 A toy example

The following gives a toy example of such a system. It is obviously not sufficient, but it will allow us to understand and reason about the system a bit.

Let us take the following two (simplified) sentences from the English Wikipedia:

"San Francisco is the cultural, commercial, and financial center of Northern California. It is the fourth-most populous city in California, after Los Angeles, San Diego and San Jose."

One advantage that is already given is that many of the entities in these two sentences are already represented as unambiguous Wikidata item identifiers:[1] San Francisco is Q62, Northern California Q1066807, California Q99, Los Angeles Q65, San Diego Q16552, and San Jose Q16553. Using identifiers instead of names the sentence can be rewritten to:

"Q62 is the cultural, commercial, and financial center of Q1066807. It is the fourth-most populous city in Q99, after Q65, Q16552, and Q16553."

These two sentences could be represented in a simple function-based abstract form as follows, using two Constructors, `Instantiation` and `Ranking`:

```
Instantiation(
  instance: San Francisco (Q62),
  class: object_with_modifier_and_of(
          object: center,
          modifier: and_modifier([cultural, commercial, financial]),
          of: Northern California (Q1066807)
        )
)
Ranking(
  subject: San Francisco (Q62),
  object: city (Q515),
  rank: 4,
  criteria: population (Q2625603),
  local_constraint: California (Q99),
  after: ordered([Los Angeles (Q65), San Diego (Q16552), San Jose (Q16553)])
)
```

Note that this does not entirely specify the result. For example, the second Constructor could be represented in English by a sentence such as *"San Francisco is the 4th-largest city in California, after Los Angeles, San Diego, and San Jose."* or *"Within California, San Francisco follows Los Angeles, San Diego and San Jose on the 4th rank on the list of cities by population."*

[1] The IDs are all assumed to refer to Wikidata items, e.g. see https://www.wikidata.org/wiki/Q62 for San Francisco

The names of the functions and their arguments are given here in English, but that is merely convenience. Both would be given by internal identifiers that then in turn could have labels and descriptions in any of the Wikipedia languages.

For each language, a renderer for every Constructor is needed. A rough sketch for English could look like this (entirely given in a convenient pseudo-syntax):

```
Instantiation:
  Instance + "is" + Class
Object_with_modifier_and_of:
  (if Modifier: Modifier) + Object + (if of: "of" + of)
Ranking:
  subject "is the" Ordinal(rank)"-largest" object "by" criteria
  "in" local_constraint", after" and_list(after)"."
```

In German, the respective renderer could look like the following:

```
Instantiation:
  Instance + "ist" + Class
Object_with_modifier_and_of:
  (if Modifier: Modifier) + Object + (if of: genitive(of))
Ranking:
  subject "ist, nach" And_list(after)", " Bestimmer_artikel(object)
  Ordinal(rank)"-größte" object "nach" criteria "in" local_constraint"."
```

Which would lead to these sentences: *"San Francisco ist das kulturelle, kommerzielle und finanzielle Zentrum Nord-Kaliforniens. Es ist, nach Los Angeles, San Diego und San Jose, die viertgrößte Stadt nach Bevölkerung in Kalifornien."*

This glosses over many language specific issues like agreement, saliency, anaphora creation for readable text, etc.

This paper is not about a concrete solution – no such solution is known at the time of writing – but about introducing the challenge of a multilingual Wikipedia, with the following questions: How would the knowledge representation look like? How would the individual articles be represented? How would the renderers be written?

3 Desiderata

The main feature of Wikipedia is that anyone can contribute to it. This is also a cornerstone of the challenge discussed in this paper.

Content must be easy to contribute – to create, refine, and change. Content will continue to constitute the largest part of Wikipedia by far, followed by the lexical knowledge and then, far behind, the renderers and the core software running the project. If trade-offs in the designs of the different systems are needed to be made, these should take into account how many contributors will be necessary for each of these parts.

Content must be editable, creatable, and maintainable in any language. No matter what language a contributor speaks, they must be able to contribute to the content of multilingual Wikipedia.

The set of available Constructors that can be used in the individual articles has to be under control of and be extensible by the community. The Constructors and their individual slots, whether these slots are required or optional, etc., have to be editable and extensible. It cannot be assumed that a full set of functions can be created a priori that will allow to capture all of Wikipedia. This also means that the system has to be able to deal with changes of Constructors – it cannot be assumed that a Constructor will be created perfect on the first try. When a Constructor changes, Content has to be transferred with as little loss as possible. All changes to Constructors have to be reversible, in order to allow for the wiki-principles that allowed Wikipedia to grow to remain applicable.

The Renderers have to be written by the community. The only way to scale the creation of the renderers to hundreds of languages is to enable the community to create and maintain them. This does not mean that every single community member must be able to write renderers: it will be a task which can only be done by contributors who dedicate some time to learn the syntax and the system. The current Wikipedia communities have shown that contributors with very different skillsets can successfully work together: some people gain deep knowledge in using templates in MediaWiki, other write bots in Java or Python to operate on a large amount of articles, others create JavaScript widgets on top of Wikipedia, and others contribute content through the VisualEditor, a WYSIWIG interface to Wikipedia.

Lexical knowledge must be easy to contribute. The renderers will have the task to translate Constructors into natural language. This will require large amounts of lexical knowledge, far larger than the renderers themselves. Fortunately, Wikidata has been recently extended in order to be able to express and maintain the lexical knowledge needed for the renderers.[2]

All pieces are changeable at all times. The system will not follow a waterfall process: it is unreasonable to expect that first all Constructors will be created, and then the required renderers will be written, then the community identifies and provides the required lexical knowledge, and finally uses it to write the content. Instead, all of these parts – as well as other parts of the infrastructure – will change and improve incrementally. The system must be able to cope with such continuous and incremental changes on all levels. On the other hand, the changes will not be entirely arbitrary: on Wikipedia there are currently many tens of thousands of templates and millions of invocations of these templates. The change to a template might require the change to all articles using the template, and often these are found to be inconsistent and articles may be broken because of that. It is OK to make certain assumption with regards to the way the different part of the system may evolve, but they have to allow for the organic growth of all the pieces.

Graceful degradation. The different languages will grow independently from each other. Because different languages will have different amounts of activity, some languages are expected to have very complete renderers, a full lexical knowledge base, and to keep up-to-date with changes to Constructors, whereas

[2] see https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

other languages will have only basic lexical knowledge, incomplete renderers, and a stale support for the existing Constructors. It is important that the system degrades gracefully, and doesn't stop rendering the whole article because of a missing lexicalization. A sentence that would be rendered in English as *"In 2013, the secretary of state was the first foreign official to visit the country since the revolution."* could degrade to *"The minister visited the country in 2013."* Parts of the Content could be marked as optional or to require other parts of the Content to be rendered, and thus allow to be dropped in case the language resources are insufficient or have become dated.

4 Unique advantages

Whereas the project sketched out here is ambitious, there are several constraints that can be taken advantage of:

- we only aim at a single genre of text, encyclopedias.
- in particular, we do not aim to represent poems, literature, dialogues, fiction, etc.
- in general, the exact surface text that is rendered is not so important, as long as it contains the necessary content and, even more important, does not introduce incorrect statements.
- in cases where the text cites other genres and relies on a high-fidelity reproduction of the text for a citation a mechanism to actually cite sources verbatim can be maed available.
- it is fine to start with extremely simple sentences and allow the community to iterate over them. In fact, this is probably the only way the project can grow. The system will likely never achieve the same expressivity as natural language, but already the ability to express simple statements and the possibility for the community to grow this ability over time is expected to make huge amounts of knowledge available to readers who previously could not access it.
- there is no need to understand and parse natural language. We merely need to generate it, which is usually considered a vastly simpler task. This means it is not needed to understand the entirety of a language, but merely to be able to generate a small subset of possible utterances. Whereas it would be neat to have a parser to help with the task of initially entering a text, this is no requirement as they are alternatives for the contributor to create and maintain the content.
- with Wikipedia, the unique opportunity to rely on a large number of volunteer contributors is available. Whereas it is obvious that the system should be designed in such a way that it reduces the effort that is necessary to be performed by human contributors, it can – and has to – rely on human contributions and not on machine intelligence to solve all problems.
- finally, the baseline is very low. We are not competing with comprehensive Wikipedias with a large number of active contributors, but we want to make knowledge available in languages that currently have only a small number of articles, many of which are out of date and incomplete.

5 Particular challenges

5.1 Why an *abstract* language

One question might be: why use an abstract language at all, and not just English, or a controlled version of English? If it could just be English, then it would be possible to start with the English Wikipedia – which would be a great way to kickstart the project.

Let us take a simple example sentence in order to illustrate some of the problems of using a specific natural language such as English:

I saw my uncle's car at the river.

This sentence is not translatable into a number of languages without additional information. Here are a few examples:

- *I saw...* can not be translated to a language like Croatian without knowing the gender of the speaker. If the speaker is female, one says *vidila*, for a male speaker one says *vidio*. In turn, it is almost impossible to add this information to the English sentence without sounding terribly contrived.
- *... uncle...* can not be translated to a language like Uzbek without knowing if it is the uncle on your mother's or your father's side. Even though this can be given explicitly in the English sentence, it would sound weirdly specific: *I saw my uncle from the mother's side's car at the river.*
- *... river.* can not be translated to French without further information: if the river flows into the ocean, then it is a *fleuve*, but if it does not, it is a *rivière*. Again, translating this sentence from French without losing that information would lead to a very unusual sentence.

In order for the sentence to contain all information needed for these three translations it may be stated like the following:

I, a male, saw my uncle on the mother's side's car at the river not flowing to the ocean.

Such a sentence would be highly unusual and be quite different from the above version of the sentence. This should demonstrate a few of the problems with choosing a natural language in order to store the content.

5.2 Saliency and information subsumption

The previous example demonstrates that rendering a sentence in different languages often yields surface texts with slightly different information content. The English, French, Uzbek and Croatian versions of the above sentence all contain slightly different information. It follows that the abstract syntax needs to capture not only all the information needed for rendering into *any* of the supported languages, but also that it needs to explicate whether some information is mandatory to be present in *every* rendered surface text, or if it is optional and only included for the benefit of a subset of the languages. There is a difference between the necessary information that a rendered sentence has to carry from the sufficient information that the abstract syntax aims to convey.

So one possibility for the above sentence could be sketched as follows (note that this example is very close to the grammar instead of the semantics – the actual constructs in the multilingual Wikipedia will probably be higher level. But this illustrates the point in this section more concisely):

```
experiencing :
  time : past
  mode : seeing
  agent :
    type : pronoun
    person : 1
    number : singular
   *gender : male
  patient :
    type : determined-object
    object : car
    possessor : determined_object
      object : uncle .. uncle-on-the-mothers-side
      possessor :
        type : pronoun
        person : 1
        number : singular
       *gender: male
    location:
      type : determined-object
      object : river .. river-that-doesnt-flow-into-ocean
```

In this sketch, optional information is marked with a `*`, and the `..` notation on the value describes a space of possible values for which the most salient term can be chosen by the renderer.

As the number of languages grow, more and more information will be required. Sometimes refactoring will be required: a translation into Chinese would consider the literal translation for car to be overly specific and prefer just to use vehicle. So when adding Chinese one could refactor

`object : car`

to

`object : vehicle .. car`

and each language could decide whether to use its word for vehicle, for car, or anything in between based on the salience and other considerations (e.g. style or domain).

Such refactorings would lend itself to bot-based mass edits.

6 Other approaches

In this section we discuss other approaches aiming solve the challenges on the way toward achieving Wikipedia's mission of allowing everyone to share in the sum of all knowledge.

6.1 Organic growth

The most extensively used approach is to grow the communities for each language independently, and hope that all languages will eventually have enough volunteers to create and maintain the encyclopedia. This has worked in a small number of languages, and many projects and quite a large amount of funding is geared toward the goal of strengthening the individual language communities and organically grow the smaller Wikipedia language editions.

To put the challenge into perspective: currently English Wikipedia has about 30,000 active editors. If all existing language editions were to have a similar number of active editors, there would be around 9 Million active editors - currently, there are about 70,000, so that would mean an increase by more than two orders of magnitude. Also, some languages Wikipedia is available in do not have 30,000 speakers: Manx, Mirandese, or Lower Sorbian are examples of languages with fewer speakers than English Wikipedia has active editors.

In short, it is not only unlikely but partially impossible to achieve the Wikipedia mission through organic growth.

6.2 Machine translation

Another widely used approach – mostly by readers, much less by contributors – is the use of automatic translation services such as Google Translate. Google Translate currently supports about 100 languages, about a third of the languages Wikipedia supports. Also the quality of these translations can vary widely – and achieving the quality a reader expects from an encyclopedia is challenging.

Unfortunately, the quality of translations often correlates with the availability of content in the given language: languages that already have a lot of content also have better translations. This is an inherent problem with the way machine translation works currently. Further breakthroughs in machine translation are required to overcome this bottleneck and are currently active areas of research.

A promising approach is the Content Translation Framework.[3] It allows to take an article from one Wikipedia language edition, and use it to start an article in another language edition. Machine translation helps with the initial text for the article, which then can be corrected and refined by the editor. The tool has shown a very promising uptake, with thousands of translations made available.

6.3 Wikidata

Wikidata is a structured multilingual knowledge base. The content of Wikidata is available in all Wikipedia languages (given that the labels are translated), but compared to natural language the expressivity of Wikidata is extremely limited. It is good to keep simple statements, mostly the kind of knowledge available in the so-called Wikipedia infoboxes which are on the right hand side of many articles: Berlin is located in Germany, Douglas Adams was born in

[3] https://www.mediawiki.org/wiki/Content_translation

Cambridge, Moby Dick was written by Herman Melville. But it could not express how photosynthesis works, the reasons of World War II, or Kant's categorical imperative. Wikidata is fast growing, and has more than 550 million statements and an active community of volunteers.

Extending Wikidata's expressivity to capture much more of Wikipedia's content would be extremely challenging and stretch the boundaries of knowledge representation far beyond the state of the art. Once this is achieved it is unclear whether the user interface could actually support that expressivity.

Another issue is that Wikidata is centered on a single subject per page, whereas most Wikipedia articles mention many other subjects while describing the main subject of an article. Furthermore, a lot of the content of Wikipedia is redundant, and this redundancy is essential for the ease of human understanding of the content. Wikidata has the population of San Francisco, but it doesn't explicitly state that it is the fourth-largest city by population in California. This would need to be deduced from the knowledge base as a whole.

7 Summary

The multilingual Wikipedia is a clearly defined goal for a challenging problem. Some of the challenges have been outlined, as well as desiderata and constraints, but it is unknown what the best solution is. When discussing this problem with others, it is often marked upon how ambitious and hard the problem is – which is particularly surprising when these people work on issues such as artificial intelligence, natural language understanding, or machine translation. It seems that the challenges inherent in AI and MT are not only necessarily harder than the challenges that would be encountered when realizing a multilingual Wikipedia, but they would be a true superset: everything needed to realize a multilingual Wikipedia will also be needed for AI.

It seems that the vision described here is achievable without the need for a major breakthrough in our current knowledge. This is a call for developers and the research communities to answer the challenges posed by the multilingual Wikipedia and demonstrate how the current state of the art in natural language generation, knowledge representation, and collaborative systems can complement each other to create a novel system that will enable everyone to share in the sum of all knowledge.

Acknowledgments. Many people helped develop the ideas presented here. Special thanks go to Markus Krötzsch, Lydia Pintscher, Daniel Kinzler, Philipp Cimiano, Elena Simperl, Erin Van Liemt, Hadar Shemtov, Dan Brickley, Jiang Bian, Barak Turovsky, Enrique Alfonseca, Scott Roy, Jamie Taylor, Colin H. Evans, Nancy Chang, Eric Altendorf, David Huynh, Macduff Hughes, Javier Snaider, Alex MacBride, Tatiana Libman, Daniel Keysers, Nikola Momchev, Nathan Scales, Nathanael Schärli, Michael Ellsworth, Russell Lee-Goldman, Praveen Paritosh, and Olivier Bousquet. I apologize to those I forgot to mention.

This is a widely extended version of a paper published at the ISWC 2018 Blue Sky track.